Slovene Twitter Analytics

Nikola Ljubešić[*†], Darja Fišer[‡*]

[*] Dept. of Knowledge Technologies, Jožef Stefan Institute
[†] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
[‡] Dept. of Translation, Faculty of Arts, University of Ljubljana

CMC conference 2016, 27th Sep 2016

Task

Twitter

- Large volumes of metadata-rich CMC data
- Data easily obtainable
- Reasonable terms of service

Task

Twitter

- Large volumes of metadata-rich CMC data
- Data easily obtainable
- Reasonable terms of service

Data analytics

- Explorative analysis of all / many variables available
- Very popular recently ("data science" buzzword) as
  - Most datasets were not produced by researchers, so variables were not controlled
  - Many datasets are crossreferenced – immense number of variables

Related work

- Rios and Lin (2013) analyse annual tweeting dynamics around the world – identifying interesting cultural differences
- Scheffler and Kyba (2016) investigate morning routine of German Twitter users – social norms of working life

Related work

- Rios and Lin (2013) analyse annual tweeting dynamics around the world – identifying interesting cultural differences
- Scheffler and Kyba (2016) investigate morning routine of German Twitter users – social norms of working life
- Bamman et al. (2012) analyse dependence of gender and
  - language standardness (women more predominant)
  - communication style (men informative, women involved)
  - vocabulary (women more distinct features)
- Arakawa et al. (2014) discriminate between organisations and private users, most others discriminate between more user types

Dataset

JANES corpus

- Tweets, fora, blogs, news comments, Wikipedia talk pages
- 7.5 million tweets, 107 million tokens posted by ∼9,000 users between June 2013 and January 2016

Dataset

JANES corpus

- Tweets, fora, blogs, news comments, Wikipedia talk pages
- 7.5 million tweets, 107 million tokens posted by ∼9,000 users between June 2013 and January 2016

Data collection

- Collected with TweetCat (Ljubešić et al., 2014) – based on Search Twitter API (`https://github.com/nljubesi/tweetcat`)
- Part of an emerging toolkit
  - TweetGeo for collecting geo-encoded tweets published in an area
  - TweetPub for preparing your data for publishing

Variables of interest

Metadata-based

- timestamp of publishing
- was the tweet retweeted
- was the tweet favourited

Variables of interest

Metadata-based

- timestamp of publishing
- was the tweet retweeted
- was the tweet favourited

Text-based

- text standardness (Ljubešić et al., 2015)
- sentiment (Fišer et al., 2016)

Variables of interest

Metadata-based

- timestamp of publishing
- was the tweet retweeted
- was the tweet favourited

Text-based

- text standardness (Ljubešić et al., 2015)
- sentiment (Fišer et al., 2016)

User-level

- gender of the user – male or female
- account type – private or corporate

Overview

Posting dynamics

- Daily and weekly
- Dependence on the gender and account type variables

Overview

Posting dynamics

- Daily and weekly
- Dependence on the gender and account type variables

Retweets and favorites

- Dependence on the gender and account type variables

Overview

Posting dynamics

- Daily and weekly
- Dependence on the gender and account type variables

Retweets and favorites

- Dependence on the gender and account type variables

Language standardness

- Dependence on the gender and account type variables
- Daily dynamics

Overview

Posting dynamics

- Daily and weekly
- Dependence on the gender and account type variables

Retweets and favorites

- Dependence on the gender and account type variables

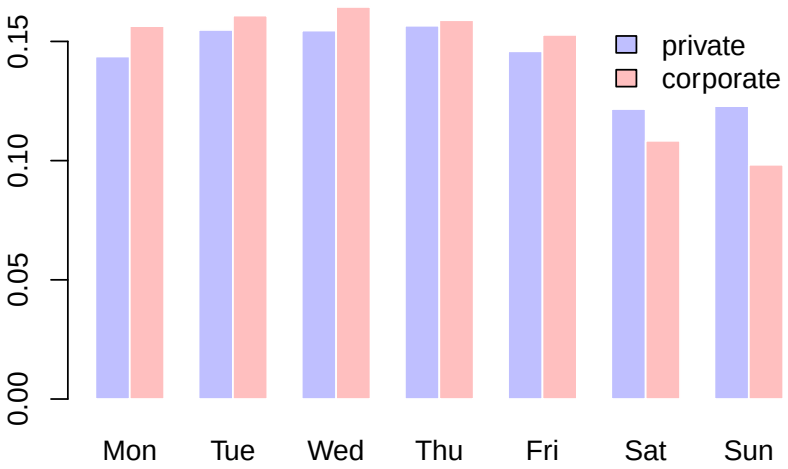
Language standardness

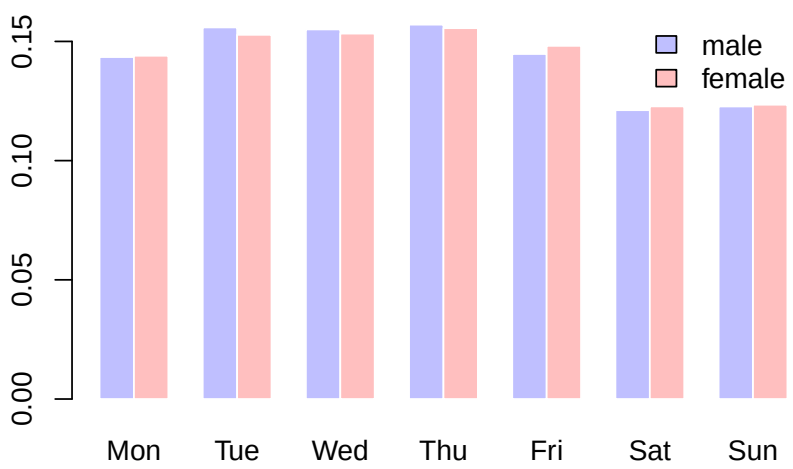
- Dependence on the gender and account type variables
- Daily dynamics

Sentiment

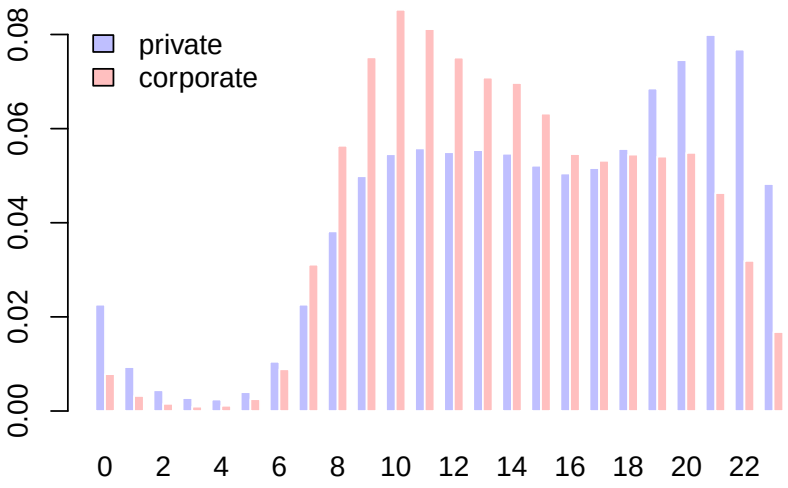
- Dependence on the gender and account type variables
- Weekly dynamics
- Dependence on the standardness variable

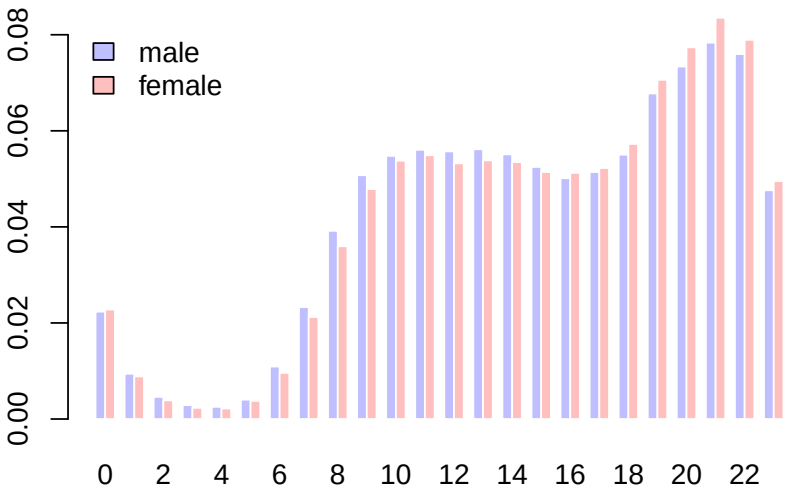
Weekly posting dynamics





Daily posting dynamics





Retweets and favorites

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

Retweets and favorites

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

- source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$

Retweets and favorites

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

- source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
- source vs. favorited $X^2(1, N = 7503200) = 80215, p < .001$

Retweets and favorites

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

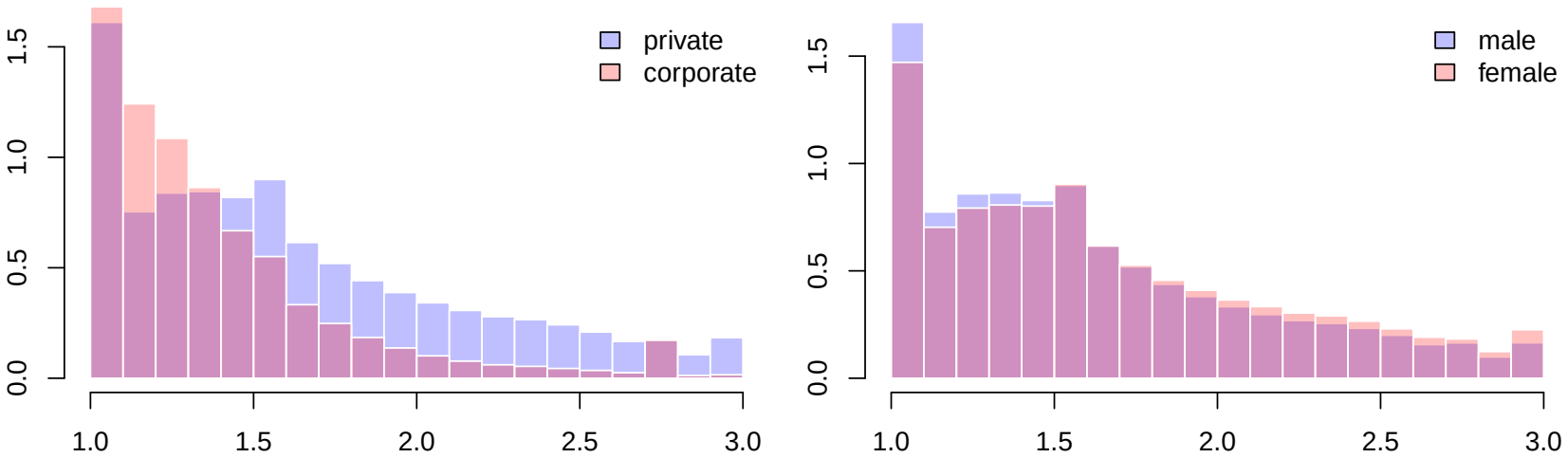
- source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
- source vs. favorited $X^2(1, N = 7503200) = 80215, p < .001$
- gender vs. retweeted $X^2(1, N = 7503200) = 11714, p < .001$

Retweets and favorites

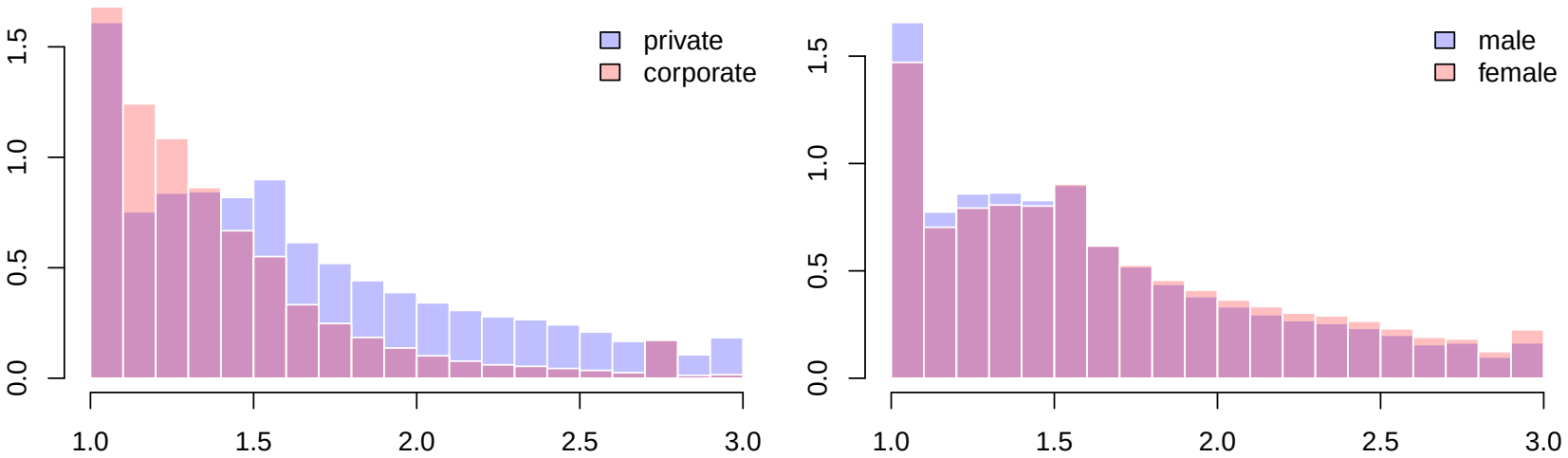
	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

- source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
- source vs. favorited $X^2(1, N = 7503200) = 80215, p < .001$
- gender vs. retweeted $X^2(1, N = 7503200) = 11714, p < .001$
- gender vs. favorited $X^2(1, N = 7503200) = 8913.4, p < .001$

Standardness vs. source and gender

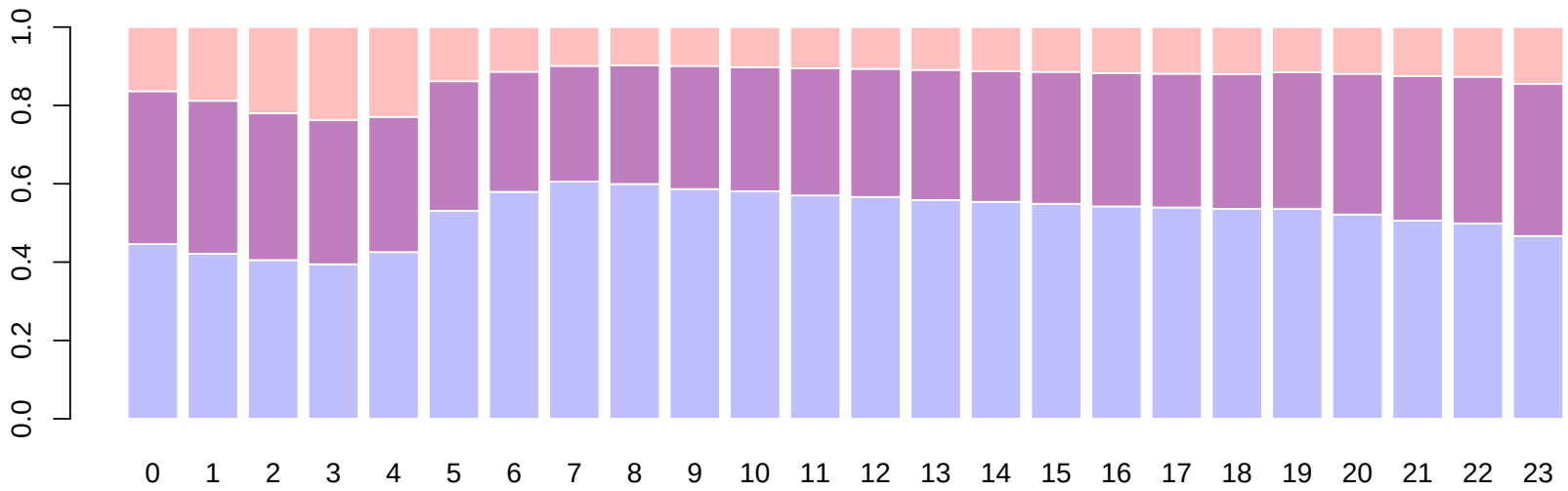


Standardness vs. source and gender

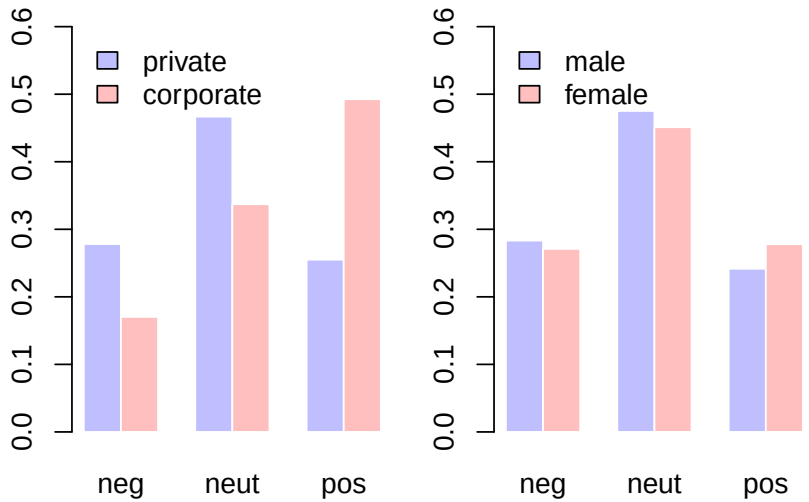


- gender vs. standardness
  $X^2(1, N = 7503200) = 9740.9, p < .001$
- 24% of female and 19.6% of male tweets very non-standard

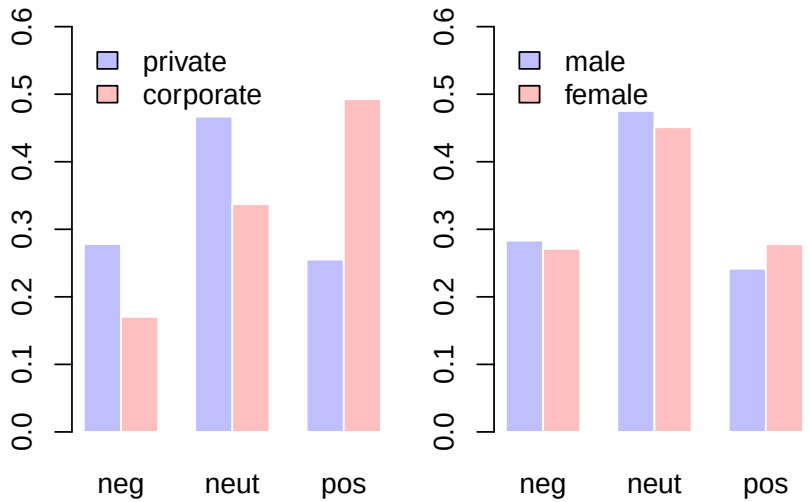
Standardness and daily dynamics



Sentiment vs. source and gender

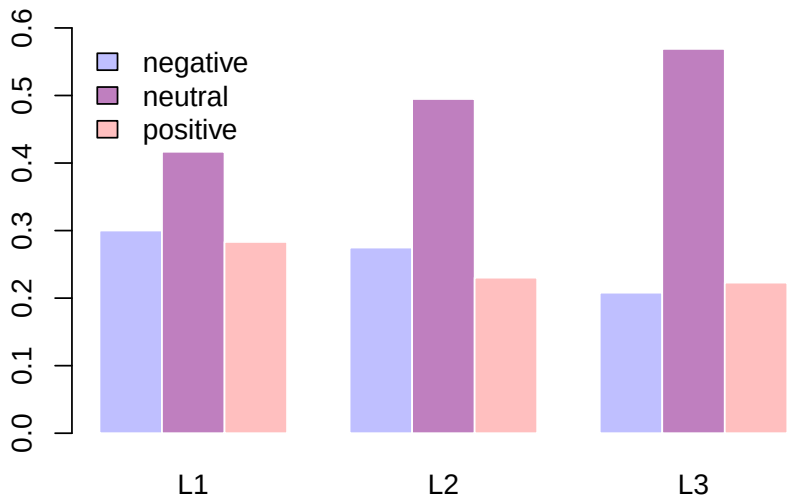


Sentiment vs. source and gender

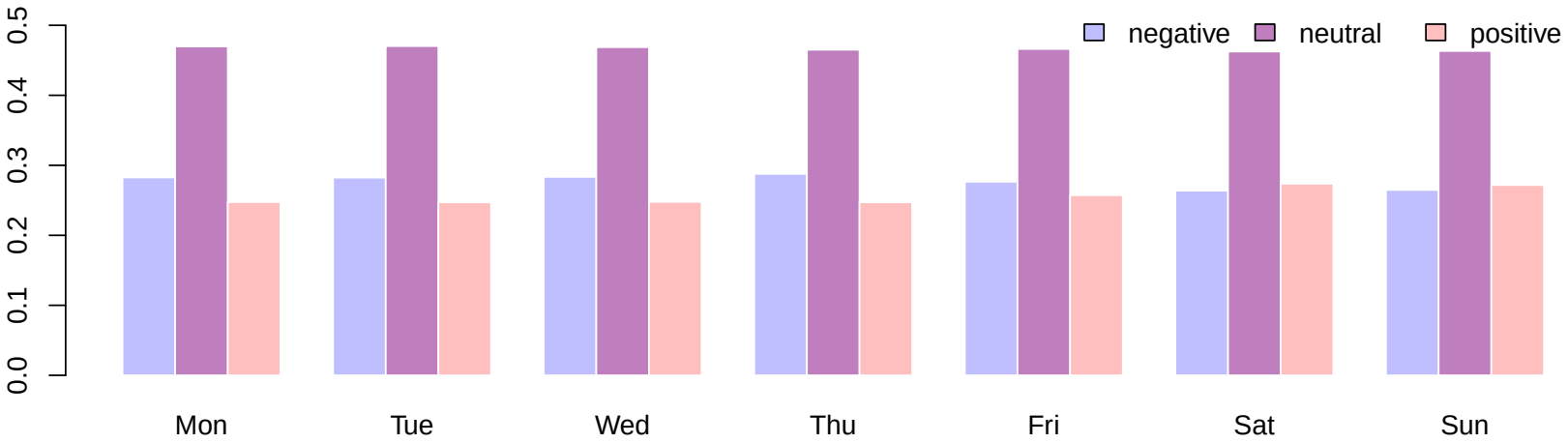


- sentiment vs. gender $X^2(1, N = 7503200) = 6179.8, p < .001$

Sentiment vs. text standardness



Weekly dynamics of sentiment



Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables

Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work

Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work
- Non-standard language used mostly in the "early" hours
- Male users tweet more during weeksdays and mornings
- Female users tweet more positive, the prevailing sentiment during weekends
- Male users use more standard language – contrary to findings of Bamman et al. (2012)

Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work
- Non-standard language used mostly in the "early" hours
- Male users tweet more during weeksdays and mornings
- Female users tweet more positive, the prevailing sentiment during weekends
- Male users use more standard language – contrary to findings of Bamman et al. (2012)
- Future work
  - More variables, more sources, more languages

Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work
- Non-standard language used mostly in the "early" hours
- Male users tweet more during weeksdays and mornings
- Female users tweet more positive, the prevailing sentiment during weekends
- Male users use more standard language – contrary to findings of Bamman et al. (2012)
- Future work
  - More variables, more sources, more languages
  - Hypotheses, in-depth analysis and comparison to related work

Slovene Twitter Analytics

Nikola Ljubešić[*†], Darja Fišer[‡*]

[*] Dept. of Knowledge Technologies, Jožef Stefan Institute
[†] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
[‡] Dept. of Translation, Faculty of Arts, University of Ljubljana

CMC conference 2016, 27th Sep 2016